

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:)
Jun IBUKI, et al.)
Serial No.: To be assigned) Group Art Unit: Unassigned
Filed: September 27, 2000) Examiner: Unassigned
For: FACT DATA UNIFYING)
METHOD AND APPARATUS)



**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN
APPLICATION IN ACCORDANCE
WITH THE REQUIREMENTS OF 37 C.F.R. §1.55**

*Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231*

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicants submit herewith a certified copy of the following foreign application:

Japanese Patent Application No. 11-310766, filed: November 1, 1999.

It is respectfully requested that the applicants be given the benefit of the foreign filing date as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,
STAAS & HALSEY LLP

Date: September 27, 2000

By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 Eleventh Street, N.W., Suite 500
Washington, D.C. 20001
(202) 434-1500

PATENT OFFICE
JAPANESE GOVERNMENT



This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: November 1, 1999

Application Number: Patent Application
No. 11-310766

Applicant(s): FUJITSU LIMITED

July 21, 2000

Commissioner,
Patent Office Kozo Oikawa

Certificate No. 2000-3055953

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT



別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1 9 9 9 年 1 1 月 1 日

出 願 番 号

Application Number:

平成 1 1 年 特 許 願 第 3 1 0 7 6 6 号

出 願 人

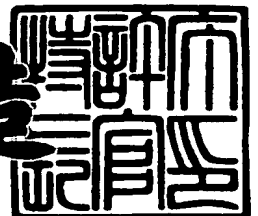
Applicant (s):

富士通株式会社

2 0 0 0 年 7 月 2 1 日

特 許 庁 長 官
Commissioner,
Patent Office

及 川 耕 造



出 証 番 号 出 証 特 2 0 0 0 - 3 0 5 5 9 5 3

【書類名】 特許願

【整理番号】 9951080

【提出日】 平成11年11月 1日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/20

【発明の名称】 事実データ統合方法および装置

【請求項の数】 11

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

 【氏名】 伊吹 潤

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

 【氏名】 落谷 亮

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

 【氏名】 西野 文人

【特許出願人】

 【識別番号】 000005223

 【氏名又は名称】 富士通株式会社

【代理人】

 【識別番号】 100100930

 【弁理士】

 【氏名又は名称】 長澤 俊一郎

 【電話番号】 03-3822-9271

【選任した代理人】

 【識別番号】 100080894

【弁理士】

【氏名又は名称】 京谷 四郎

【電話番号】 03-3823-7935

【手数料の表示】

【予納台帳番号】 024143

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9704945

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 事実データ統合方法および装置

【特許請求の範囲】

【請求項 1】 対象とする事物、属性名、属性値の 3 つ組によって規定される事実データをテキストから抽出し、

抽出された事実データについて同種のデータをまとめて、テキスト全体にわたるデータ集計を行ない、

集計されたデータ集合を走査して両立し得ない不整合データ群を検出し、

不整合データ群においてどれが正しいデータであるかを判断し、誤りデータを排除して正しい事実データの統合を行う

ことを特徴とする事実データ統合方法。

【請求項 2】 対象とする事物、属性名、属性値の 3 つ組によって規定される事実データをテキストから抽出するデータ抽出部と、

データ抽出部で抽出された事実データについて、テキスト全体にわたり同種のデータをまとめ、生起回数を集計するデータ集計部と、

データ集計部において集計されたデータ集合を走査して両立し得ない不整合データ群を検出する不整合検出部と、

不整合検出部で検出された不整合データ群の中でどれが正しいデータであるかを判断する正誤判定部と、

データ集計部において集計された正しいデータ、および、正誤判定部において正しいデータと判断されたデータを集計する最終データ集積部とを備えた

ことを特徴とする事実データ統合装置。

【請求項 3】 事実データをテキストから抽出する際にデータに信頼度を付与する信頼度付与部を設け、

データ集計部において生起回数を集計する際、集計されたデータの信頼度を個々のデータの信頼度から計算して集計結果に付与し、

正誤判定部において、上記データに付与された信頼度を利用してデータ群中の各データの正誤の判断を行なう

ことを特徴とする請求項 2 の事実データ統合装置。

【請求項 4】 上記信頼度付与部が、テキストから事実データを抽出する際に抽出の対象となったテキストの持つイベント情報の種類を判定するイベント型抽出部と、

イベント型と信頼度の対応表に基づき、イベント型から信頼度を評価する信頼度評価部とを備えている

ことを特徴とする請求項 3 の事実データ統合装置。

【請求項 5】 上記信頼度付与部が、テキスト中の抽出対象とする対象事物に対しての注目度を計算する注目度評価部と、

上記注目度に基づき、データの信頼度を評価する信頼度評価部とを備えていることを特徴とする請求項 3 の事実データ統合装置。

【請求項 6】 上記信頼度付与部が、テキストの発行者、著者等の書誌情報と該テキストに記述される各データの信頼度を対応付ける書誌情報と信頼度の対応表と、

テキスト中からデータの抽出を行なう際、上記書誌情報と信頼度の対応表を参照して該テキストの書誌情報からテキストの信頼度を評価する信頼度評価部とを備えている

ことを特徴とする請求項 3 の事実データ統合装置。

【請求項 7】 データ抽出部によって抽出する事実データに正誤フラグを付与し、正誤フラグを付加させた正誤のフラグ付きの事実データを入力として受けとり、事実データの属性名毎に特定の属性値をとるデータの正誤の期待値を計算し、書誌情報と信頼度の対応表を生成する

ことを特徴とする請求項 6 の事実データ統合装置。

【請求項 8】 対象事物、属性名と、正誤判定の際に利用する判定方法とを対応付けた属性・判定方法対応表と、

上記属性・判定方法対応表に基づき、属性名に応じた正誤判定方法を決定する判定方法決定部とを備え、

正誤判定部は、不整合データ群が入力された際、上記判定方法決定部により指定された判定方法を用いて正誤判定を行う

ことを特徴とする請求項 2, 3, 4, 5, 6 または請求項 7 の事実データ統合装

置。

【請求項 9】 データ抽出部と不整合検出部の間に誤りパターン除去部を設け、

誤りパターン除去部は、データ抽出部で抽出された事実データと、予め登録された誤りパターンとを照合することにより個々のデータ毎に正誤の判断を行ない、抽出された事実データが予め登録された誤りパターンに合致した時に誤りと判断して棄却し、問題がないとされたデータのみを不整合検出部に送ることを特徴とする請求項 2, 3, 4, 5, 6, 7 または請求項 8 の事実データ統合装置。

【請求項 1 0】 データ集計部の後にデータ統合部を設け、

データ統合部は、互いに似ているデータを統合して、一つのデータに統合した後、不整合検出部に渡すことを特徴とする請求項 2, 3, 4, 5, 6, 7 または請求項 8 の事実データ統合装置。

【請求項 1 1】 テキストから抽出された対象事物、属性名、属性値の 3 つ組によって規定される事実データを統合するデータ統合プログラムを記録した記録媒体であって、

上記データ統合プログラムは、対象とする事物、属性名、属性値の 3 つ組によって規定される事実データをテキストから抽出し、

抽出された事実データについて同種のデータをまとめて、テキスト全体にわたるデータ集計を行ない、

集計されたデータ集合を走査して両立し得ない不整合データ群を検出し、不整合データ群においてどれが正しいデータであるかを判断し、誤りデータを排除して正しい事実データの統合を行う

ことを特徴とするデータ統合プログラムを記録した記録媒体。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は、文書中の事実の記述を抽出して整合性をもつデータの集合としてデ

ータベース化したり、また事実データの矛盾点から対応する元テキストの持つ誤りの検出や訂正をする事実データ統合方法および装置に関する。

【0002】

【従来の技術】

テキスト中の情報の抽出技術としては従来から種々の方法が提案されており、例えば、新製品の情報、組織体情報等の予め決められた枠組に従ったデータの場合は、テキスト中の表現形式と抽出されるデータとの対応表を保持しておき、テキストを走査して規定された表現形式にマッチした時に対応するデータを取り出すことが行なわれている。

例えば、図20(a)に示すような対応表を保持しておき、入力テキストを走査して、同図(b)(c)に示すように「対象事物」、「属性名」、「属性値」からなる事物データを抽出する。この例の場合には、入力テキストの「C社の新社長」、「D氏に決定した」がそれぞれ対応表の*1,*2にマッチするので、同図(d)に示すように対象事物としては「C社」、属性名として「代表者」、属性値として「D氏」が抽出される。

一方、誤りの訂正技術に関しては、テキスト中に存在する表記レベルの誤りに対象を限った場合には様々な技術が既に存存している。例えば、テキスト中に存在する表現を登録しておき未登録語を指摘したり、表記の揺れの指摘などの方法が知られている。

【0003】

【発明が解決しようとする課題】

上述のようにテキストからの事実データの抽出は広く行なわれているが、見たい情報がテキスト中の一箇所からの情報だけで得られるとは限らないので一般にはテキスト全体からのデータを統合することが必要となる。

ところが、テキスト自体の含む誤り、あるいは抽出処理の誤り等によって一般には抽出されるデータ中にかなりの誤り（あるいはデータの不統一）が存在し、これらを人手でチェックして除いたり、書き換える必要があるために単純に集計することができなかった。

本発明は上記した事情を考慮してなされたものであって、テキスト中の誤った

記述や抽出処理の誤りに起因する抽出データ中の誤りやバラツキの訂正や標準化を行うことにより、適切なデータの集積を可能とすることを目的とする。

【0004】

【課題を解決するための手段】

図1は本発明の基本構成を示すブロック図である。同図において、1は対象とする事物、属性名、属性値の3つ組によって規定される事実データをテキストから抽出するデータ抽出部、2は同種のデータをまとめ、生起回数を集計するデータ集計部、3はデータ集計部において集計されたデータ集合を走査して両立し得ない不整合データ群を検出する不整合検出部、4は不整合データ群の中でどれが正しいデータであるかを判断する正誤判定部、5はデータ集計部において集計された正しいデータ、および、正誤判定部において正しいデータと判断されたデータを集計する最終データ集積部である。

また、6は事実データをテキストから抽出する際にデータに信頼度を付与する信頼度付与部、7は互いに似ているデータを統合して、一つのデータに統合するデータ統合部、8は予め登録された誤りパターンに合致した事実データを誤りとして棄却する誤りパターン除去部、9は正誤判定部における正誤判定方法を決定する判定方法決定部である。

【0005】

図1に示すように本発明においては、次のようにして前記課題を解決する。

(1) 対象とする事物、属性名、属性値の3つ組によって規定される事実データをテキストから抽出するデータ抽出部1と、データ抽出部1で抽出された事実データについて、テキスト全体にわたり同種のデータをまとめ、生起回数を集計するデータ集計部2と、データ集計部2において集計されたデータ集合を走査して両立し得ない不整合データ群を検出する不整合検出部3と、不整合検出部3で検出された不整合データ群の中でどれが正しいデータであるかを判断する正誤判定部4と、データ集計部2において集計された正しいデータ、および、正誤判定部において正しいデータと判断されたデータを集計する最終データ集積部5とを設け、抽出された事実データから誤りデータを排除して適切なデータの集積を可能とする。

(2) 上記(1)において、事実データをテキストから抽出する際にデータに信頼度を付与する信頼度付与部6を設け、データ集計部2において生起回数を集計する際、集計されたデータの信頼度を個々のデータの信頼度から計算して集計結果に付与し、正誤判定部4において、上記データに付与された信頼度を利用してデータ群中の各データの正誤の判断を行なうことにより正誤判断の精度を高める。

(3) 上記(2)において、上記信頼度付与部6を、テキストから事実データを抽出する際に抽出の対象となったテキストの持つイベント情報の種類を判定するイベント型抽出部と、イベント型と信頼度の対応表に基づき、イベント型から信頼度を評価する信頼度評価部とから構成し、正確な信頼度を付与する。

(4) 上記(2)において、上記信頼度付与部6を、テキスト中の抽出対象とする対象事物に対しての注目度を計算する注目度評価部と、上記注目度に基づき、データの信頼度を評価する信頼度評価部とから構成し、正確な信頼度を付与する。

(5) 上記(2)において、上記信頼度付与部6を、テキストの発行者、著者等の書誌情報と該テキストに記述される各データの信頼度を対応付ける書誌情報と信頼度の対応表と、テキスト中からデータの抽出を行なう際、上記書誌情報と信頼度の対応表を参照して該テキストの書誌情報からテキストの信頼度を評価する信頼度評価部とから構成し、作者、発行者等による一般的な傾向を考慮した信頼度を付与する。

(6) 上記(5)において、データ抽出部1によって抽出する事実データに正誤フラグを付与し、正誤フラグを付加させた正誤のフラグ付きの事実データを入力し、事実データの属性名毎に特定の属性値をとるデータの正誤の期待値を計算し、書誌情報と信頼度の対応表を生成することにより、属性値と信頼度の対応表を半自動的にテキストから生成する。

(7) 上記(1)～(6)において、対象事物、属性名と、正誤判定の際に利用する判定方法とを対応付けた属性・判定方法対応表と、上記属性・判定方法対応表に基づき、属性名に応じた正誤判定方法を決定する判定方法決定部とを設け、不整合データ群が入力された際、上記判定方法決定部により指定された判定方法

を用いて正誤判定部により正誤判定を行うことにより、属性に応じた柔軟な正誤判断を行なう。

(8) 上記(1)～(7)において、データ抽出部1と不整合検出部の間に誤りパターン除去部を設け、誤りパターン除去部8において、データ抽出部1で抽出された事実データと、予め登録された誤りパターンとを照合することにより個々のデータ毎に正誤の判断を行ない、抽出された事実データが予め登録された誤りパターンに合致した時に誤りと判断して棄却し、問題がないとされたデータのみを不整合検出部に送ることにより、単独で判断可能な誤りの除去を行なう。

(9) 上記(1)～(6)において、データ集計部2の後にデータ統合部7を設け、データ統合部7において、互いに似ているデータを統合して、一つのデータに統合した後に不整合検出部3に渡すことにより、同じ事物の異なる表現による揺らぎを吸収する。

【0006】

【発明の実施の形態】

以下本発明の実施の形態について説明する。

図2は本発明の事実データ統合処理を行うためのシステムの構成例を示す図である。同図において、101はCRT、液晶ディスプレイ等の表示装置、キーボード、マウス等の、文字、記号、命令等を入力するための入力装置から構成される入出力装置、102はCPU、103はROM、RAM等から構成されるメモリ、104はプログラム、データ等を記憶する外部記憶装置、105はフロッピーディスク、MO、CD-ROM等の可搬型記憶媒体にアクセスしてデータの読み出し／書き込みを行う媒体読み取り装置、106は電話回線を使用してデータ通信をするためのモデム、LAN等のネットワークを使用してデータ通信を行うためのネットワークカード等を含む通信インタフェースである。

外部記憶装置104には本発明の事実データ統合処理を行うプログラム、事実データを抽出するテキストデータが格納されており、また、事実データ統合処理を行った結果得られた統合データ等が格納される。

【0007】

図3は本発明の第1の実施例の機能ブロック図であり、同図により本発明の第

1 の実施例について説明する。

図 3 において、1 1 はテキスト中の事実データに関する記述を解析し、事実データとして抽出するデータ抽出部、1 2 はデータ抽出部 1 1 において抽出された事実データの内、同じデータを一つにまとめて各事実データの生起回数を計数するデータ集計部、1 3 はデータ不整合検出部であり、テキスト中から抽出された事実データ集合中における不整合（例えば、両立できないような事実データの組み合わせ）を捜し出す。1 4 はデータ不整合検出部 1 3 で検出された不整合データのどれが正しくどれが誤っているかを判断する正誤判定部、1 5 は正しいと検証されたデータを集積して提示する最終データ集積部である。

【0 0 0 8】

図 3 において、テキストデータが入力されると、データ抽出部 1 2 では、前記従来例で説明したように、テキスト中の記述を解析し、事実データとして抽出する。

図 4（a）は、前記図 2 0（a）に示した対応表を用い、テキスト中から該対応表に規定された表現形式の事実データを抽出した場合におけるデータ抽出部 1 2 の出力例であり、前記した対応表によれば図 4（a）に示すように対象事物（A 社、F 社、…、H 社）、属性名（代表者、…、所在）、属性値（B、G、…、C 国）からなる事実データが抽出される。

データ集計部 1 2 では、上記事実データをソートして同じデータをまとめ、各事実データの生起回数を計数する。図 4（b）は同図（a）に示した事実データについてのデータ集計部 1 2 の出力例を示す図であり、同図に示すように「対象事物」、「属性名」、「属性値」と、それらが一致する事実データの生起回数が出力される。

【0 0 0 9】

不整合検出部 1 3 は、事実データ集合中での不整合データを検出する。そのため例えば次のような処理を行う。

- i) データ集合中で全ての対象事物に対して以下の操作を繰り返す。
- ii) 選択した対象事物のもつ全ての属性名について以下の操作を繰り返す。
- iii) 同じ属性名に対応する属性値が複数存在すれば、そのデータ群を不整合デ

ータ群として出力し、それ以外は整合データとして出力する。

【0 0 1 0】

図 4 (c) は不整合検出部 1 3 において不整合データとして検出された不整合データ例を示す図であり、同図に示すようにデータ集計部 1 2 で集計された事実データの内、対象事物「A 社」、属性名「代表」について、B 氏と D 氏の 2 種類の値があるので、属性値「B」と「D」が不整合データとして検出され正誤判定部 1 4 に送られる。また、データ集計部 1 2 で集計された残りのデータは、整合データとして最終データ集積部 1 5 に送られる。

正誤判定部 1 4 では不整合データについてどれが正しくどれが誤っているかを判断する。

これについては次のように様々なアルゴリズムが考えられる。

- i) 群中の最大生起回数をもつデータを正しいと判断し、他を誤りとする。
- ii) 特定の閾値以上の生起回数をもつデータを正しいと判断し他を誤りとする。

【0 0 1 1】

図 4 (d) は正誤判定部 1 4 の出力例を示す図であり、同図は図 4 (c) の不整合データについて、上記 i) のアルゴリズムにより正誤判定をした場合の出力例を示している。

不整合データとして検出された対象事物「A 社」、属性名「代表」の属性値「B」、「D」の内、属性値「B」の生起回数が 2 件、「D」の生起回数が 1 件であるので、この例では、図 4 (d) に示すように属性値「B」が「正」として採用され、属性値「D」が誤りとして棄却される。

【0 0 1 2】

最終データ集積部 1 5 では、上記不整合検出部 1 3 から整合データとして送られてきたデータおよび正誤判定部 1 4 で正しいデータとして判定されたデータを集積して提示する。図 4 (e) は最終データ集積部 1 5 の出力例を示す図であり、同図に示すように、データ集計部 1 2 で集計されたデータの内、不整合データ検出部 1 3 から整合データとして送られてきたデータおよび正誤判定部 1 4 で正しいとして判定されたデータが正しいデータとして出力される。

【0 0 1 3】

図 5 は本実施例の処理を示すフローチャートであり、同図により上記処理を説明する。

図 5 において、ステップ S 1 において、入力されたテキストデータの事実データに関する記述を解析して事実データとして抽出し、例えば前記図 4 (a) に示したような事実データを得る。

ステップ S 2 において、抽出された事実データを対象事物、属性名、属性値についてソートし、ソートしたデータをカウントする。その結果、前記図 4 (b) の示したデータが得られる。

【0014】

ステップ S 3 において、ソートされた対象事物を一つ取り出す。ステップ S 4 において、取り出した対象事物のなかの一つの属性名を選択し、ステップ S 5 においてその整合性をチェックする。そして、例えば前記図 4 (c) に示したような不整合データが検出された場合には、ステップ S 6 に行き、前記①、②に示したアルゴリズムにより不整合データの正誤判定を行い、誤ったデータを棄却する。また、データが整合している場合には、ステップ S 7 において、整合データを集積する。

ステップ S 8 において、属性名についての整合性チェックが尽くされたかを判定し、尽くされていない場合には、ステップ S 4 に戻り上記処理を繰り返す。また、属性名の整合性チェックが尽くされた場合には、ステップ S 9 において、対象事物についての整合性チェックが尽くされたかを判定し、尽くされていない場合には、ステップ S 3 に戻り上記処理を繰り返す。また、対象事物についての整合性チェックが尽くされた場合には処理を終了する。

【0015】

図 6 は本発明の第 2 の実施例の機能ブロック図であり、本実施例は、第 1 の実施例において、信頼度付与部を設け、テキストデータの信頼度を付与し信頼度に基づき正誤判断を行うようにしたものである。

同図において、データ抽出部 1 1 は前記したように、テキスト中の事実データに関する記述を解析し、データとして抽出する。また、信頼度付与部 1 6 はデータ抽出の対象となるテキストのもつ情報を利用して抽出したデータの信頼度の評

価を行う。

【0016】

具体的な評価方法としては例えば次のような手法を用いることができる。

①イベント型による信頼度の評価

部分テキストからイベント型を抽出しこれにより部分テキストの信頼度を評価する。

②注目度による信頼度の評価

対象事物のテキスト中における注目度に着目し、信頼度の評価を行う。

③書誌情報による信頼度の評価

テキストのもつ書誌情報（著者、発行媒体等）によって信頼度を評価する。例えば、テキストが新聞記事の場合にはその新聞が一般紙か、スポーツ紙か等のニュースソースによって信頼度を評価する。

【0017】

次に、データ集計部 12 では信頼度付のデータのデータ集計を行なうため、個々の信頼度からデータ集計としての信頼度を計算する。

このアルゴリズムとしては次のようなものが考えられる。

i) 個々のデータの信頼度の内で最大のものをデータ集計の信頼度とする。

ii) 個々のデータの信頼度の平均をデータ集計の信頼度とする。

正誤判断部 15 においてはデータ集計のもつ信頼度、生起回数を元にしてどのデータが正しいかの判断を行なう。このアルゴリズムとしては次のようなものが考えられる。

i) 個々のデータの信頼度の内で最大のものを正しいとし残りを全て誤りとする。

ii) 信頼度の閾値を定め、特定の値以上の信頼度をもつデータを正しいとする。

【0018】

図 7 は図 6 に示した信頼度付与部 16 の第 1 の内部構成例を示す図であり、この例は上記①のイベント型により信頼度を評価する場合の構成を示している。

図 7 において、11 は前記したテキストから事物データを抽出するデータ抽出部であり、データ抽出部 11 は前述したようにテキスト中の事実データに関する記述を解析し、データとして抽出する。例えば、原文が図 8 (a) に示すように

「A社の代表に…」、「A社のD社長が…」、「A社はBを…」の場合、同図に示すように「対象事物」として、「A社」、「属性名」として「代表」、「製品」、「属性値」として「B氏」、「D社長」、「B」が抽出される。

【0019】

16は信頼度付与部であり、信頼度付与部16におけるイベント型抽出部16aは原文から図8(b)に示すようなキーワード群を抽出し、図8(c)に示すキーワード・イベント対応表16cを参照して、テキスト中に存在するキーワードが表中の値とマッチした場合に対応するイベント型をもつと判断する。その結果、図9(e)に示すように抽出対象となった部分テキストからイベント型が抽出される。

信頼度評価部16bでは図8(d)に示すイベント型・信頼度対応表16dを参照して、図9(f)に示すようにイベント型により、事実データのもつべき信頼度を評価する。また、イベント型に対応しないものは、defaultとして信頼度を例えば0.5とする。

以上のようにして信頼度を付与することにより、例えば死亡記事は特に人物データに入念なチェックがかかるため人事異動等に関する記事より信頼度が高いなどの知識を利用して、正確な信頼度の評価を行うことができる。

【0020】

図10は上記信頼度付与部の第2の内部構成例を示す図であり、この例は前記②の注目度により信頼度を評価する場合の構成を示している。

図10において、11は前記したように事物データ自体の抽出を行なう事物データ抽出部、16は信頼度付与部であり、16eは抽出対象となった事物のテキスト中における注目度の評価を行なう注目度評価部、16fは注目度により信頼度を評価する信頼度評価部である。

【0021】

注目度評価部16eにおける注目度の評価手法としては、次のようなアルゴリズムが考えられる。

i)対象事物の直後につく助詞を調べ、かかり助詞「は」、「も」等がついた事物の注目度を最も高い値とし、それ以外の場合注目度を低い値とする。

例えば、図 11 (a) に示すように、注目度を上記かかり助詞がついた主語は 0.8、目的語は 0.5、その他の要素は 0.4 のように定め、図 11 (b) に示すように、原文中の事物データが上記主語であるか、目的語であるか、その他の要素であるかを判定し、それに応じて注目度を設定する。

ii) 対象事物のテキスト中の位置（先頭から何番目の単語であるか）を数え、それを位置と注目度の対応表を利用して注目度を評価する。

例えば、図 11 (c) に示すように、単語の位置と注目度の対応表を用い、原文中の事物データの位置に応じて、注目度を設定する。

信頼度評価部 16 f は、上記のようにして抽出された注目度を利用して、事実データの持つべき信頼度を計算する。基本的には、注目度の高い事物に対する信頼度が上がるように評価アルゴリズムを設定する。例えば、図 11 (d) に示すように、注目度が閾値 α より大きいかな否かを調べ、それに応じて信頼度を付与する。

以上のように、係り助詞や対象事物のテキスト中の位置等の情報を利用して注目されている事物の信頼度を上げることにより正確な正誤の判断ができる。

【0022】

図 12 は上記信頼度付与部の第 3 の内部構成例を示す図であり、この例は前記 ③の書誌情報により信頼度を評価する場合の構成を示している。

図 12 において、11 は前記したように事物データ自体の抽出を行なうデータ抽出部、16 は信頼度付与部であり、信頼度付与部 16 の信頼度評価部 16 g は、テキストの持つ書誌情報（発行元、著者等）を入力として受け、書誌情報・信頼度の対応表 16 h を利用して事実データのもつべき信頼度を調べる。

例えば、テキストの信頼度を発行元により評価し、信頼性の高い発行元であるかな否かにより対応した信頼度を付与する。

【0023】

以下、図 13 の具体例により説明する。例えば、図 13 (a) に示すように原文テキストの記述に対応した書誌情報（発行元）がそれぞれ「A新聞社」、「B新聞社」、「C通信社」であり、書誌情報・信頼度の対応表 16 h が例えば図 13 (b) に示すように「A新聞社」、「B新聞社」、「C通信社」についてそれぞ

れ信頼度が0.6、0.8、0.9として設定されている場合、信頼度付与部16は上記書誌情報・信頼度の対応表16hにより各テキストに信頼度を付与し、データ抽出部11から出力される事物データには、図13(c)に示すようにそのニュースソースに応じた信頼度が付与される。

【0024】

前記図2に示したデータ集計部12では前記したi)またはii)のアルゴリズムにより上記信頼度付き事実データを集計し、不整合検出部13に渡す。不整合検出部13では、図13(c)の事実データの内、A社の代表が「B」と「D」で不整合であるので、図13(d)に示すように上記A社の代表の「B」と「D」を不整合データとして、信頼度を付して正誤判定部14へ出力する。

正誤判定部14では、例えば前記i)またはii)のアルゴリズムを用いて、正誤判定を行う。例えば、前記i)の「群内で信頼度の最も高いものを正として選択し、他のデータを誤りとする」を用いて正誤判定を行う場合には、図13(e)に示すようにA社の代表の「B」を誤りとして棄却し、「D」を正としてデータ集積部15へ出力する。その結果、データ集積部15からは図13(f)に示すデータが出力される。

【0025】

図14は、図12に示した信頼度付与部における書誌情報・信頼度の対応表16hを生成するための構成例を示している。

図14において、書誌情報属性走査部17には正誤のフラグ付きの事実データが入力される。正誤のフラグは、事実データが正しいか否かを示すフラグであり、例えば人手で予め付与しておいたりあるいは他のシステムで自動的に付与するようにしてもよい。

書誌情報属性走査部17は書誌情報等の各属性の値毎にデータ全体を探索し、その属性の持つ事実データを抽出する。例えば、前記した「A新聞社」、「B新聞社」、「C通信社」、…等のニュースソースについての信頼度を得る場合には、上記新聞社、通信社毎にデータ全体を探索し、正誤フラグが付されている事実データを抽出する。

【0026】

信頼度評価部 1 8 は、上記書誌情報属性走査部 1 7 において抽出された事実データについて、正誤フラグを元にデータの正解率を計算し、各書誌情報毎の信頼度を得る。これにより、例えば、上記した「A新聞社」、「B新聞社」、「C通信社」、…のそれぞれの信頼度を得ることができる。

データ登録部 1 9 は上記信頼度評価部 1 8 において求めた信頼度を書誌情報・信頼度の対応表 1 6 h に登録し、データベース化する。

上記のようにして書誌情報・信頼度の対応表 1 6 h を生成することにより、人手で対応表へデータを登録する手間を省くことができる。

【0 0 2 7】

図 1 5 は本発明の第 3 の実施例を示す図であり、本実施例は、前記図 3、図 6 において、判定方法決定部 2 0 を設けて、正誤判定部 1 5 における判定方法を決定するようにしたものであり、その他の構成は前記図 3、図 6 と同じである。

図 1 5 において、不整合検出部 1 4 に不整合データ群が入力されるとまず判定方法決定部 2 0 において、事実データの対象事物、属性名を調べる。ついで、判定方法決定部 2 0 は、属性・判定方法対応表 2 1 を参照し、正誤判断のための方法を決定する。属性・判定方法対応表 2 1 は、対象事物、属性名とそれに応じた判定方法が予め登録されている。

【0 0 2 8】

例えば、属性・判定方法対応表 2 1 には、属性名が「部長」のように該当する人間が複数いる可能性のある場合には、「一定の閾値以上の信頼度をもつデータは全て正しいとする」というような第 1 の判定方法が登録され、また、属性名が「社長」のように 1 人しかいない場合には、「信頼度が最も高いデータのみを正しいとする」というような第 2 の判定方法が登録されており、判定方法決定部 2 0 は、属性名が「部長」の場合には上記第 1 の判定方法を指定し、属性名が「社長」の場合には上記第 2 の判定方法を指定する。

正誤判定部 1 5 は判定方法決定部 2 0 において指定された正誤判定方法によりデータ群内の各データの正誤の判定を行う。

以上のようにして正誤判定を行うことにより、会社の部長のように複数の値をもつことが可能なデータと社長のようにユニークな値をもつことしか許されない

データに対して独自の正誤判断をすることが可能となる。

【0029】

図16は、本発明の第4の実施例を示す図であり、本実施例は、前記第1、第2の実施例において、誤りパターン除去部22を設け単独のデータとして誤りであると判断されるデータを棄却するようにしたものであり、その他の構成は前記図3、図6と同じである。

図16において、データ集計部12と不整合検出部13の間に誤りパターン除去部22が設けられており、誤りパターン除去部22は、データ集計部12からデータが与えられたとき、誤りパターンデータベース23を参照して、単独のデータとして誤りであると判断されるデータを棄却する。

【0030】

図17は図16に示した誤りパターン除去部22における誤りパターンの判断例を示す図である。この例では、電話番号における誤りパターンとして、頭が0でない数字が来るものを規定して誤りを検出する例を示している。

例えば、データ抽出部11により抽出されたデータが図17(a)に示すようにA社とB社の電話番号である場合、誤りパターン除去部22では、誤りパターンデータベース23を参照し、電話番号についての誤りパターンと比較する。

ここでは、誤りパターンデータベース23に図17(b)に示す誤りパターンが登録されていたとする。ここで、図17(b)は0で始まらない電話番号は誤りであることを正規表現で表記したものである。

誤りパターン除去部22で図17(a)に示す電話番号と、図17(b)に示す誤りパターンを比較すると、B社の電話番号「119-0003」は0以外の数字で始まっているので、誤りであると判定され、図17(c)に示すようにB社の電話番号が棄却される。

【0031】

図18は本発明の第5の実施例を示す図であり、本実施例はデータ統合部を設け、類似の属性値を持つデータの統合を行なうことにより、表記の揺れに対処するようにしたものであり、その他の構成は前記図6と同じである。

図18において、データ集計部12と不整合検出部13の間にデータ統合部2

4 が設けられており、データ統合部 2 4 は、データ揺れデータベース 2 5 を参照して類似の属性値を持つデータの統合を行なう。これにより、表記の揺れによって実際には大量に生起しているのに各表記に対してはあまり多くの生起例がないように見られ正誤の判断を誤る場合に対処することができる。

【0 0 3 2】

図 1 9 は本実施例の処理例である。データ抽出部 1 1 において図 1 9 (a) に示すようなデータが抽出されると、データ統合部 2 4 では、類似の値を持つデータの統合を行なう。

この例ではデータ揺れデータベース 2 5 に、人名データの統合条件として「姓名が示された人名と、姓のみの人名は類似データとして統合可能」という条件が設定されているとする。データ統合部 2 4 は、データ揺れデータベース 2 5 を参照して上記条件により、A 社の属性名「代表」の属性値として「山田一郎」をもつデータと「山田」をもつデータを統合する。その結果図 1 9 (b) に示すように A 社の属性名「代表」のデータが統合され、データの頻度は両者の件数の和とされる。

【0 0 3 3】

データ統合部 2 4 において上記のようにデータの統合が行われると、正誤判定部 1 4 での正誤判定においては、上記統合された頻度により正誤判定を行う。例えば、前記した「群中の最大生起回数をもつデータを正しいと判断し、他を誤りとする」というアルゴリズムに正誤判定を行う場合には、図 1 9 (c) に示すように A 社の「代表者」について「山田太郎」が正とされ、「鈴木太郎」が誤りとされる。

この例の場合、「山田一郎」、「山田」のそれぞれの生起回数より「鈴木太郎」の生起回数の方が多いのでデータ統合を行わない場合には、「鈴木太郎」が正とされることとなるが、上記のようなデータ統合を行うことにより、正しい正誤の判断を行なうことが可能となる。

【0 0 3 4】

【発明の効果】

以上説明したように、本発明においては、以下の効果を得ることができる。

(1) 事実データをテキストから抽出し、抽出された事実データについて同種のデータをまとめて、テキスト全体にわたるデータ集計を行ない、集計されたデータ集合を走査して両立し得ない不整合データ群を検出し、不整合データ群においてどれが正しいデータであるかを判断し、誤りデータを排除して正しい事実データの統合を行うようにしたので、テキスト中の誤った記述や抽出処理の誤りに起因する抽出データ中の誤りやバラツキに対して、誤り部分を排除して、適切なデータの集積を行なうことができる。

(2) 事実データをテキストから抽出する際にデータに信頼度を付与し、信頼度を利用してデータ群中の各データの正誤の判断を行なうことにより正誤判断の精度を高めることができる。

(3) 属性名に応じて正誤判定の際に使用する判定方法を指定し、該判定方法により正誤判定を行うことにより、属性に応じた柔軟な正誤判断を行なうことができる。

(4) 抽出した事実データと予め登録された誤りパターンとを照合し、抽出された事実データが予め登録された誤りパターンに合致した時に誤りと判断して棄却することにより、単独で判断可能な誤りの除去を行なうことができる。

(5) 互いに似ているデータを統合して、一つのデータに統合した後に不整合検出を行うことにより、同じ事物の異なる表現による揺らぎを吸収することができる。

【図面の簡単な説明】

【図 1】

本発明の基本構成を示すブロック図である。

【図 2】

事実データ統合処理を行うためのシステムの構成例を示す図である。

【図 3】

本発明の第 1 の実施例を示す図である。

【図 4】

第 1 の実施例の処理例を示す図である。

【図 5】

本発明の第 1 の実施例の処理を示すフローチャートである。

【図 6】

本発明の第 2 の実施例の機能ブロック図である。

【図 7】

信頼度付与部の第 1 の内部構成例を示す図である。

【図 8】

図 7 に示す信頼度付与部における処理例（1）を示す図である。

【図 9】

図 7 に示す信頼度付与部における処理例（2）を示す図である。

【図 1 0】

信頼度付与部の第 2 の内部構成例を示す図である。

【図 1 1】

図 1 0 に示す信頼度付与部における処理例を示す図である。

【図 1 2】

信頼度付与部の第 3 の内部構成例を示す図である。

【図 1 3】

図 1 2 に示す信頼度付与部における処理例を示す図である。

【図 1 4】

書誌情報・信頼度の対応表を生成するための構成例を示す図である。

【図 1 5】

本発明の第 3 の実施例を示す図である。

【図 1 6】

本発明の第 4 の実施例を示す図である。

【図 1 7】

第 4 の実施例における誤りパターンの判断例を示す図である。

【図 1 8】

本発明の第 5 の実施例を示す図である。

【図 1 9】

第 5 の実施例の処理例を示す図である。

【図 2 0】

テキスト中の情報の抽出方法を説明する図である。

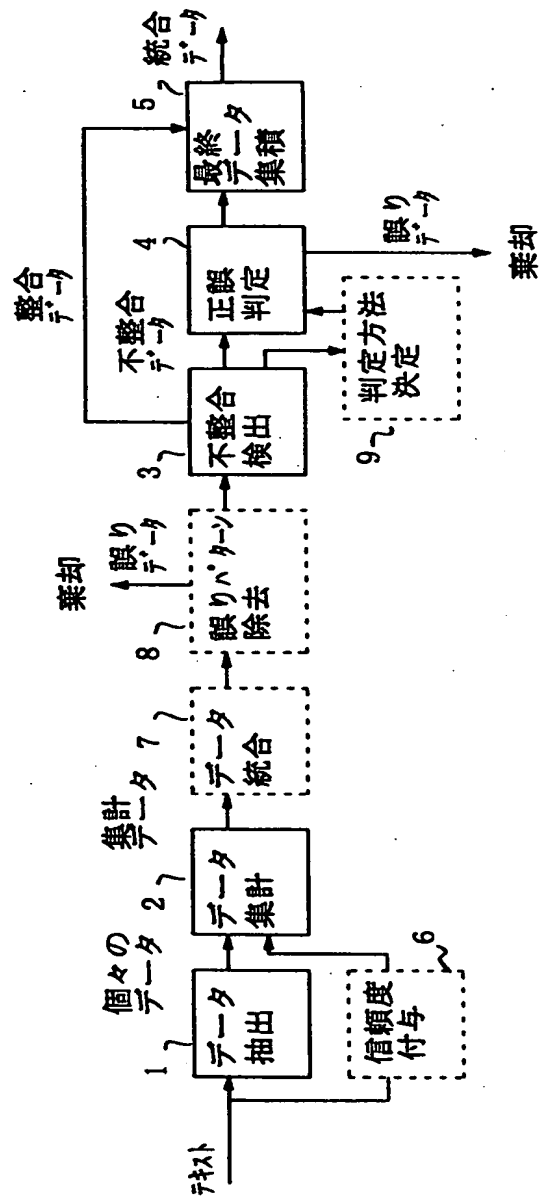
【符号の説明】

- 1, 1 1 データ抽出部
- 2, 1 2 データ集計部
- 3, 1 3 不整合検出部
- 4, 1 4 正誤判定部
- 5, 1 5 最終データ集積部
- 6, 1 6 信頼度付与部
- 7 データ統合部
- 8 誤りパターン除去部
- 9 判定方法決定部
- 1 7 書誌情報属性走査部
- 1 8 信頼度評価部
- 1 9 データ登録部
- 2 0 判定方法決定部
- 2 1 属性・判定方法対応表
- 2 2 誤りパターン除去部
- 2 3 誤りパターンデータベース
- 2 4 データ統合部
- 2 5 データ揺れデータベース

【書類名】 図面

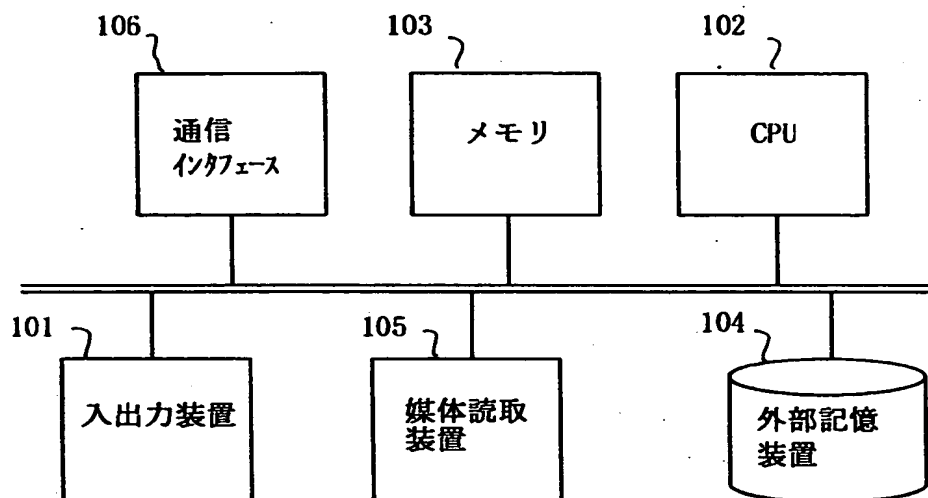
【図 1】

本発明の基本構成を示すブロック図



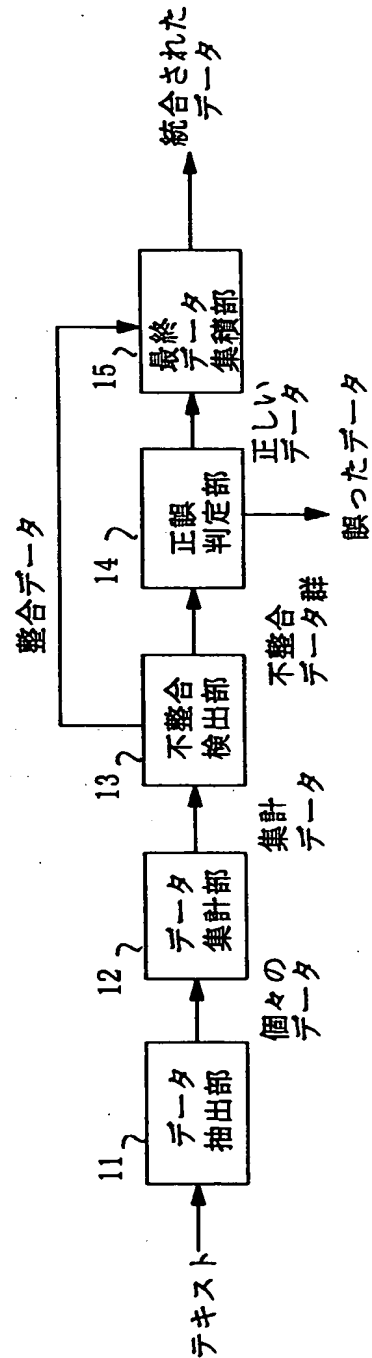
【図 2】

本発明の実施例の処理を行うためのシステムの構成例を示す図



【図 3】

本発明の第1の実施例を示す図



【図 4】

第 1 の実施例の処理例を示す図

(a) データ抽出部の出力例

対象事物	属性名	属性値
A社	代表	B
F社	代表	G
A社	代表	B
A社	代表	D
H社	所在	C国

(b) データ集計部の出力例

対象事物	属性名	属性値	生起回数
A社	代表	B	2
F社	代表	G	1
A社	代表	D	1
H社	所在	C国	1

(c) 不整合検出部の出力例

対象事物	属性名	属性値	生起回数
A社	代表	B	2
A社	代表	D	1

(d) 正誤判定部の出力例

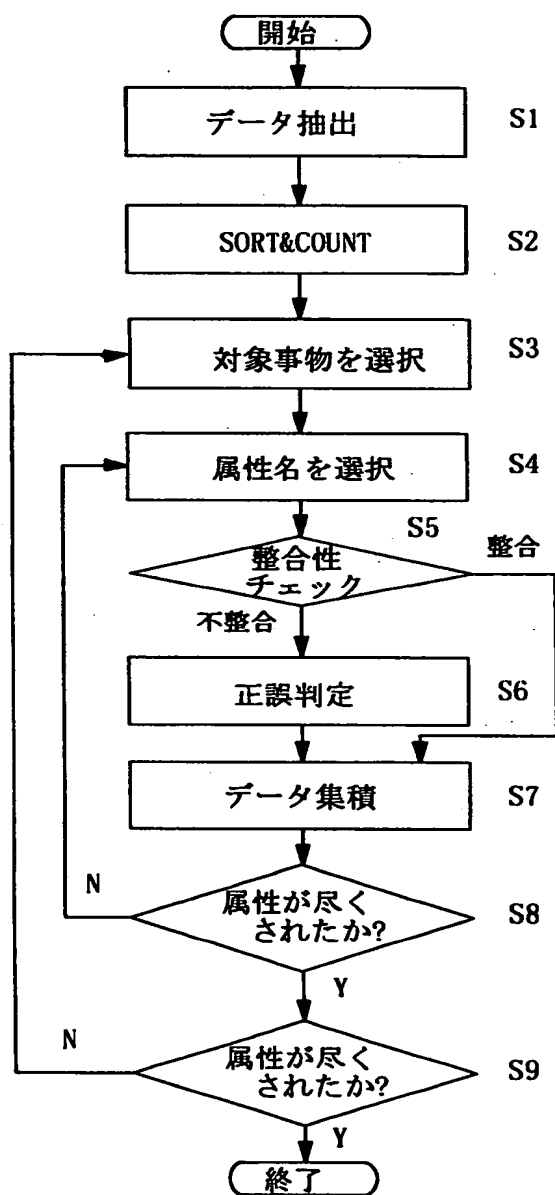
データ			正誤フラグ
A社	代表	B	正
A社	代表	D	誤 → 棄却

(e) 最終データ集積部の出力例

対象事物	属性名	属性値
A社	代表	B
F社	代表	G
H社	所在	C国

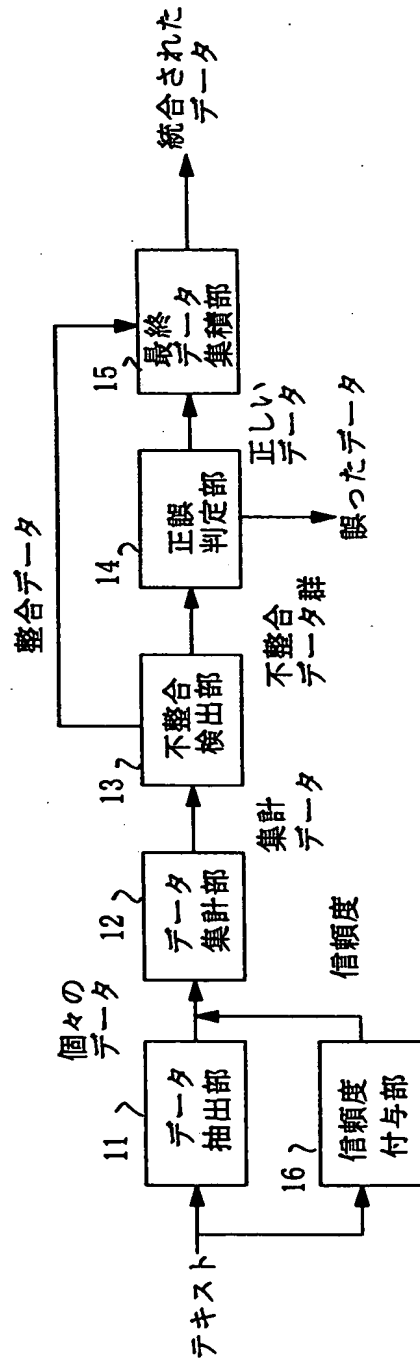
【図 5】

本発明の第 1 の実施例の処理を示すフローチャート



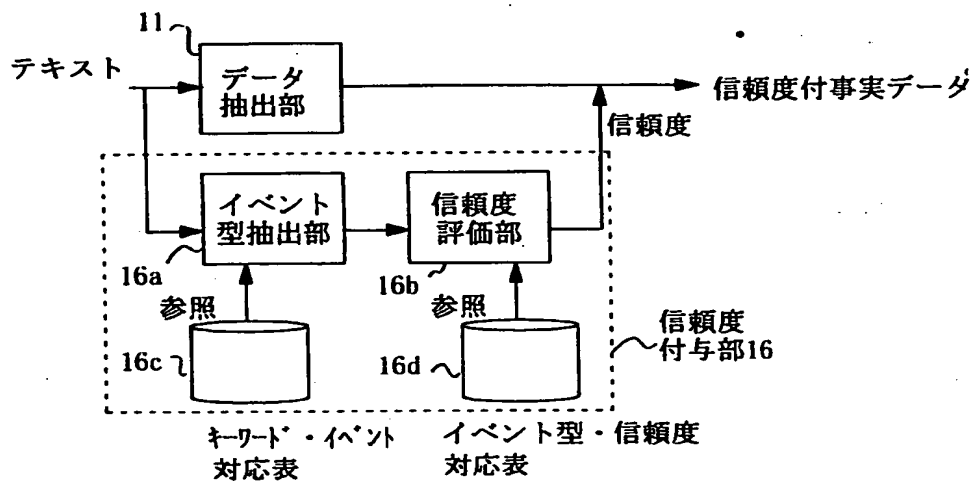
【図 6】

本発明の第 2 の実施例の機能ブロック図



【図 7】

信頼度付与部の第1の内部構成例を示す図



【図 8】

図 7 に示す信頼度付与部における処理例 (1) を示す図

(a) 原文と抽出データ例

原文	対象事物	属性名	属性値
A社の代表にB氏が就任	A社	代表	B氏
A社のD社長が逝去	A社	代表	D社長
A社はBを発売	A社	製品	B

(b) 原文とキーワードの対応例

原文	抽出されたキーワード
A社の代表にB氏が就任	A社, A社, B氏, 就任
A社のD社長が逝去	A社, D社長, 逝去
A社はBを発売	A社, B, 発売

(c) キーワード、イベント型対応表例

キーワード	イベント型	信頼度
就任、解任	人事移動	0.8
死去	死亡記事	0.9

(d) イベント型・信頼度対応表例

イベント型	信頼度
人事移動	0.8
死亡記事	0.9
default	0.5

【図 9】

図 7 に示す信頼度付与部における処理例 (2) を示す図

(e) テキストのイベント型判断例

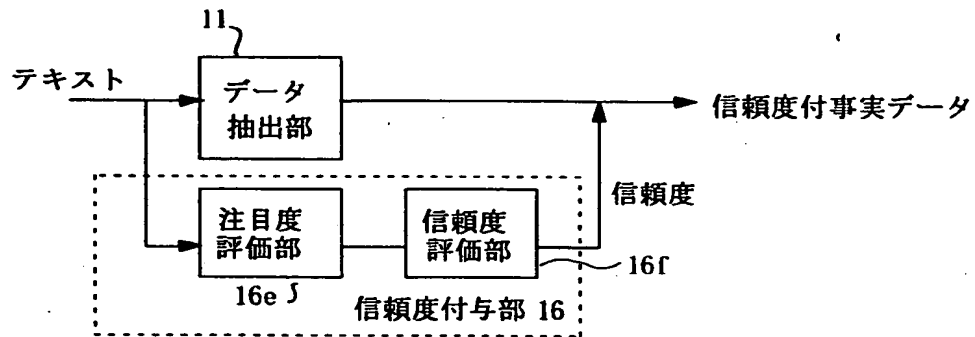
原文	抽出されたキーワード	イベント型
A社の代表にB氏が就任	A社、代表、B氏、就任	人事移動
A社のD社長が逝去	A社、D社長、逝去	死亡記事
A社はBを発売	A社、B、発売	default

(f) 各データに付与された信頼度例

原文	イベント型	信頼度
A社の代表にB氏が就任	人事移動	0. 8
A社のD社長が逝去	死亡記事	0. 9
A社はBを発売	default	0. 5

【図 1 0】

信頼度付与部の第2の内部構成例を示す図



【図 1 1】

図 1 0 に示す信頼度付与部における処理例を示す図

(a) 注目度の評価方法の例

主語	0. 8
目的語	0. 5
その他の要素	0. 4

(b) 原文中の事物に付与された注目度の例
(主語、目的語の順で注目度が高いとして
注目度を設定する例)

原文	<u>A社のB社長は新製品群を発表</u>		
	↓	↓	↓
注目度	0. 4	0. 8	0. 5
<hr/>			
原文	<u>A大臣はB社長と懇談</u>		
	↓	↓	
注目度	0. 8	0. 4	

(c) 単語の位置と注目度対応表例

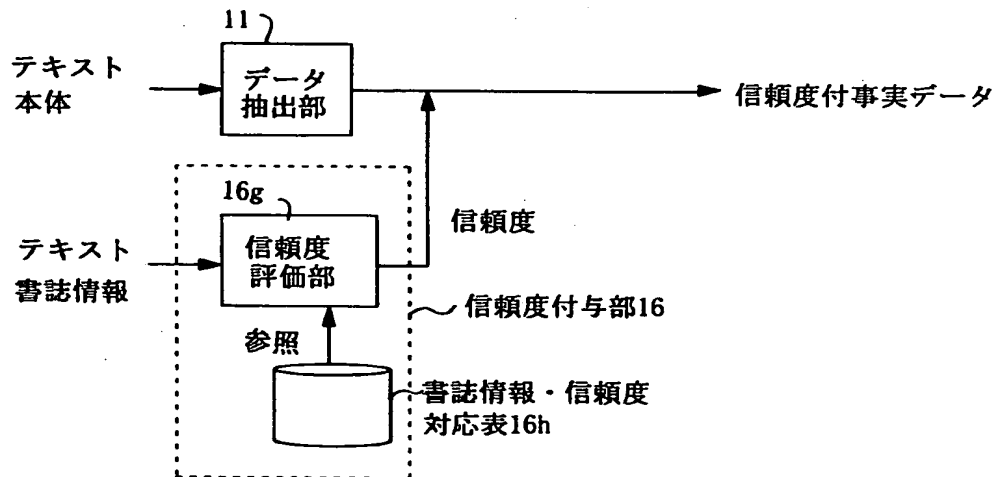
位置 < 5	注目度 = 5 - 位置
位置 ≥ 5	注目度 = 0

(d) 信頼度評価アルゴリズム例

注目度 > α	信頼度 = 0. 9
注目度 ≤ α	信頼度 = 0. 7

【図 1 2】

信頼度付与部の第 3 の内部構成例を示す図



【図 1 3】

図 1 2 に示す信頼度付与部における処理例を示す図

(a) 原テキスト中の記述と書誌情報の例

テキスト	媒体
A社のB社長は…	A新聞社
A社（代表者D）	B新聞社
A社（本社：C県E市）	C通信社

(b) 書誌情報と信頼性の対応表の例

媒体名	信頼度
A新聞社	0. 6
B新聞社	0. 8
C通信社	0. 9

(c) データ抽出部の出力例

対象事物	属性名	属性値	信頼度
A社	代表	B	0. 6
A社	代表	D	0. 8
H社	所在	C国	0. 9

(d) 不整合検出部の出力例

対象事物	属性名	属性値	信頼度
A社	代表	B	0. 6
A社	代表	D	0. 8

(e) 正誤判定部の判断例

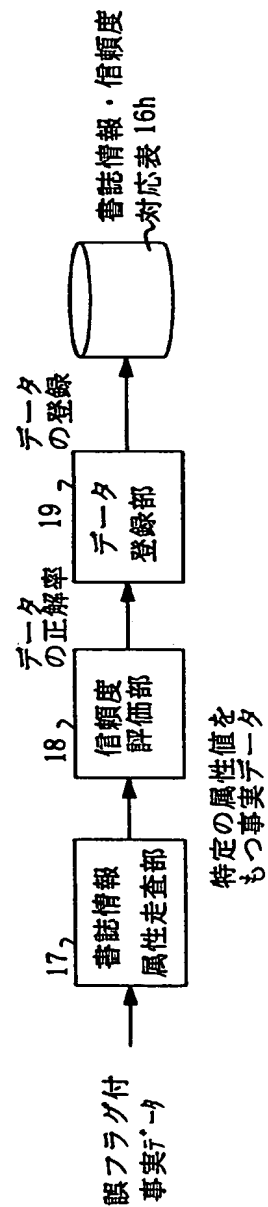
対象事物	属性名	属性値	信頼度	正誤
A社	代表	B	0. 6	誤
A社	代表	D	0. 8	正

(f) データ集積部の出力例

対象事物	属性名	属性値
A社	代表	B

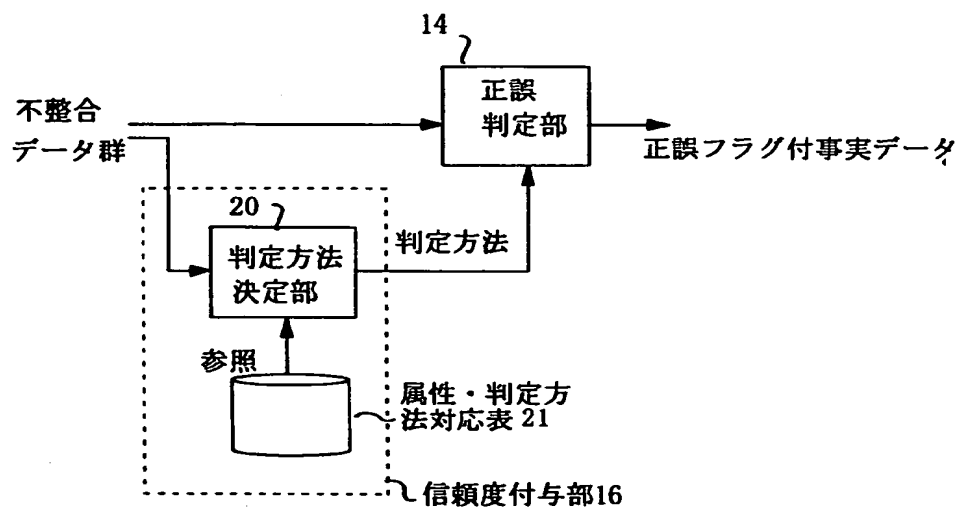
【図 1 4】

書誌情報・信頼度の対応表を生成するための構成例を示す図



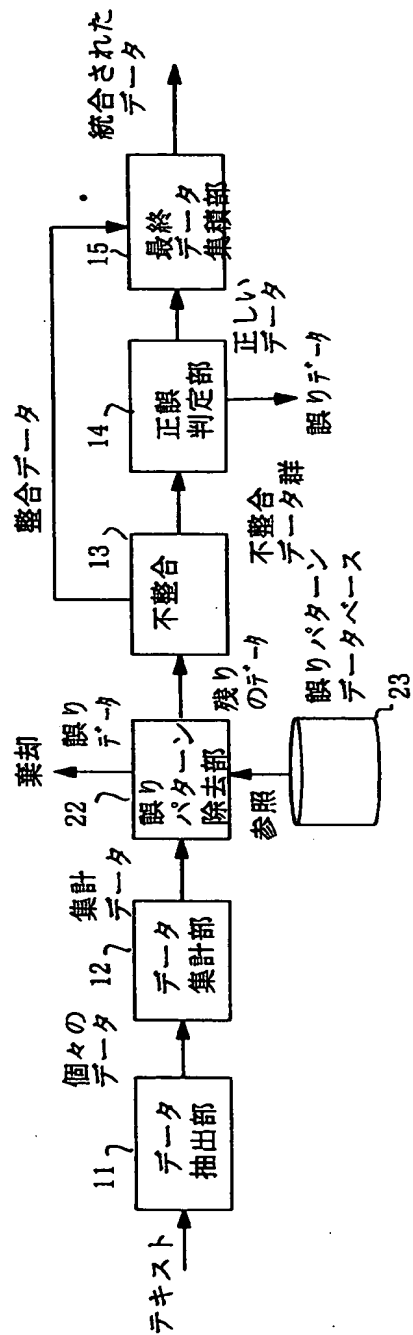
【図 1 5】

本発明の第 3 の実施例を示す図



【図 1 6】

本発明の第 4 の実施例を示す図



【図 1 7】

第 4 の実施例における誤りパターンの判断例を示す図

(a) 抽出データ例

A社	電話番号	03-356-7098
B社	電話番号	119-0003

(b) 誤りパターン例

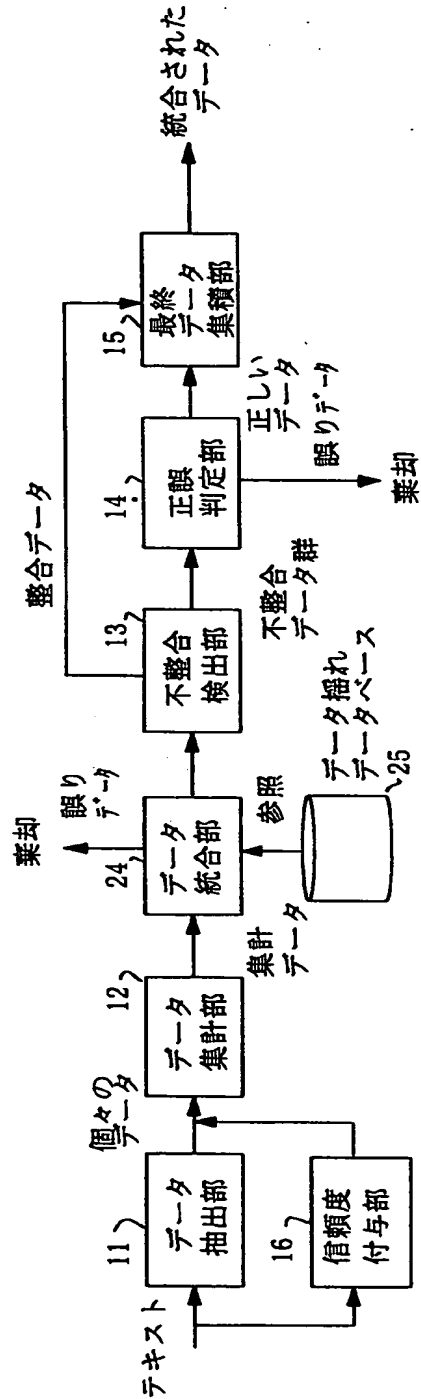
属性名	正規表現	意味
電話番号	^(^0)[0-9]+	0で始まらない数字

(c) 正誤判断例

データ			正誤
A社	電話番号	03-356-7098	正
B社	電話番号	119-0003	誤

【図 1 8】

本発明の第 5 の実施例を示す図



【図 1 9】

第 5 の実施例の処理例を示す図

(a) 抽出データの例

対象事物	属性名	属性値	頻度
A社	代表者	山田一郎	20件
A社	代表者	山田	20件
A社	代表者	鈴木太郎	30件

(b) データ統合処理結果の例

対象事物	属性名	属性値	頻度
A社	代表者	山田一郎 (山田の20件追加)	40件
A社	代表者	鈴木太郎	30件

(c) 正誤判断例

対象事物	属性名	属性値	頻度	正誤
A社	代表者	山田一郎	40件	正
A社	代表者	鈴木太郎	30件	誤

【図 2 0】

テキスト中の情報の抽出方法を説明する図

(a) 対応表の例

【表現形式】	* 1 の社長に * 2 が決定
--------	------------------

(b) 抽出データ

対象事物	属性名	属性値
.....
* 1	代表者	* 2

(c) 処理例

【入力テキスト】
A社とB社の合併によって設立されるC社の新社長に
(* 1 にマッチ)
D氏が決定した。
(* 2 にマッチ)

(d) 抽出されるデータ

対象事物	属性名	属性値
.....
C社	代表者	D氏

【書類名】 要約書

【要約】

【課題】 テキスト中の誤った記述や抽出処理の誤りに起因する抽出データ中の誤りやバラツキの誤り部分を排除して、適切なデータの集積を行なうこと。

【解決手段】 データ抽出部 1 により、対象とする事物、属性名、属性値の 3 つ組によって規定される事実データをテキストから抽出し、抽出された事実データについて、データ集計部 2 でテキスト全体にわたり同種のデータをまとめ、生起回数を集計する。不整合検出部 3 はデータ集計部 2 において集計されたデータ集合を走査して両立し得ない不整合データ群を検出し、正誤判定部 4 において、不整合データ群の中でどれが正しいデータであるかを判断する。最小データ集積部 5 は、正しいデータを集積して出力する。また、事実データをテキストから抽出する際にデータに信頼度を付与し、データに付与された信頼度を利用してデータ群中の各データの正誤の判断を行なうこともできる。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000005223]

1. 変更年月日	1996年 3月26日
[変更理由]	住所変更
住 所	神奈川県川崎市中原区上小田中4丁目1番1号
氏 名	富士通株式会社